(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－306205

(43)公開日　平成11年(1999)11月５日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/403	３３０Ｃ
			15/40	３７０Ａ

審査請求　有　請求項の数７　ＦＤ　（全 21 頁）

(21)出願番号　特願平10－129485

(22)出願日　平成10年(1998)４月23日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　島津　秀雄
東京都港区芝五丁目７番１号　日本電気株式会社内

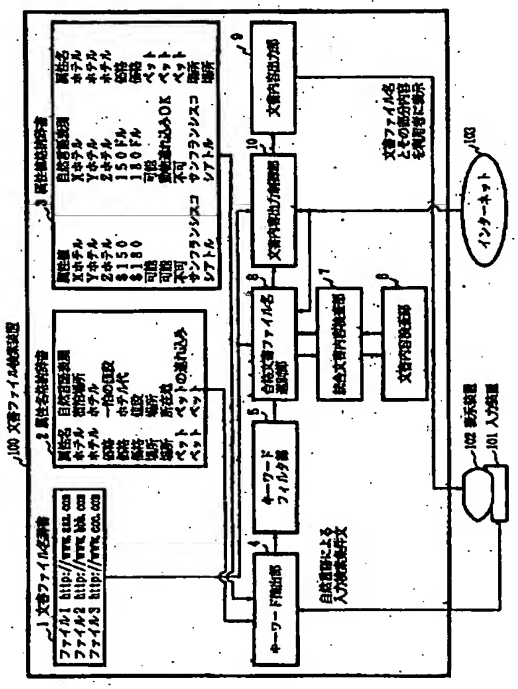
(74)代理人　弁理士　境　廣巳

(54)【発明の名称】　文書ファイル検索装置及びプログラムを記録した機械読み取り可能な記録媒体

(57)【要約】

【課題】　WWW のホームページに対する自然言語による検索問い合わせを実現する。

【解決手段】　検索対象文書ファイルであるWWW のホームページをXML で記述する。検索条件文が入力されるとキーワード抽出部４は、属性名を表現する自然言語表現はその属性名を含む属性名インデックスに、属性値を表現する自然言語表現はその属性値と属性名との対を含む属性値インデックスに変換する。キーワードフィルタ部５は、変換後のインデックス列中で、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する箇所の属性名インデックスを削除する。文書内容検査部６は、検索対象文書ファイル中に、変換後のインデックス列中の全ての属性値インデックスの属性名と属性値との対に対応するタグが存在するか否かを調べ、存在する場合、文書内容出力部９が変換後のインデックス列中の属性名インデックスの属性名を持つタグの属性値を検索して出力する。



1

【特許請求の範囲】

【請求項１】　属性の属性名とその属性の属性値との対を内蔵する文書ファイルを検索対象文書ファイルとし、検索対象文書ファイルから、利用者が自然言語で指定した検索条件に適合する部分を検索する文書ファイル検索装置において、

自然言語で表現した検索要求文を先頭から順に探査し、属性名を表現する自然言語表現に対してはその属性名を属性名インデックスとして出力し、属性値を表現する自然言語表現に対してはその属性値と属性名との対を属性値インデックスとして出力することを順次行うキーワード抽出部と、

前記キーワード抽出部の出力を入力して先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部と、

検索対象文書ファイル中に、前記キーワードフィルタ部から出力された全ての属性値インデックスの属性名と属性値との対が内蔵されているか否かを調べ、内蔵されている場合、前記キーワードフィルタ部から出力された属性名インデックスの属性名に対応する属性値を検索対象文書ファイルから検索して出力する検索手段とを備えた文書ファイル検索装置。

【請求項２】　検索対象文書ファイル中に存在する属性名について、属性名とその属性名を表現する自然言語表現との対を格納しておく属性名格納辞書と、

検索対象文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組を格納しておく属性値格納辞書とを備え、

前記キーワード抽出部は、自然言語で表現した検索要求文を先頭から順に探査し、属性名格納辞書を参照して、属性名を表現する自然言語表現が含まれていたら、その自然言語表現と対である属性名を属性名インデックスとして出力し、属性値格納辞書を参照して、属性値を表現する自然言語表現が含まれていたら、その自然言語表現と３つ組である属性値と属性名との対の集合を属性値インデックスとして出力する構成を有することを特徴とする請求項１記載の文書ファイル検索装置。

【請求項３】　文書中に書かれた意味を表現する属性名のついたタグとその属性の属性値との対を内蔵する文書ファイルを検索対象文書ファイルとし、検索対象文書ファイルから、利用者が自然言語で指定した検索条件に適合する部分を検索する文書ファイル検索装置において、

検索対象文書ファイル中に存在する属性名について、属性名とその属性名を表現する自然言語表現との対の集合を格納しておく属性名格納辞書と、

検索対象文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組の集合を格納しておく属性値格納辞書と、

自然言語で表現した検索要求文を先頭から順に探査し、属性名格納辞書を参照して、属性名を表現する自然言語表現が含まれていたら、その自然言語表現と対である属性名を属性名インデックスとして出力し、属性値格納辞書を参照して、属性値を表現する自然言語表現が含まれていたら、その自然言語表現と３つ組である属性値と属性名との対を属性値インデックスとして出力するキーワード抽出部と、

キーワード抽出部の出力を入力し、先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部と、

検索対象文書ファイル中に、前記キーワードフィルタ部から出力された全ての属性値インデックスの属性名と属性値との対に対応するタグの対が内蔵されているか否かを調べ、内蔵されている場合、前記キーワードフィルタ部から出力された属性名インデックスの属性名を持つタグの属性値を検索対象文書ファイルから検索して出力する検索手段とを備えた文書ファイル検索装置。

【請求項４】　文書中に書かれた意味を表現する属性名のついたタグとその属性の値との対を複数個内蔵する文書ファイルの集合から、利用者が自然言語で指定した検索条件を満足する文書ファイルを選択してその適合する部分を表示する文書ファイル検索装置において、

検索対象となるすべての文書ファイルの名前と存在位置とを格納する文書ファイル名辞書と、

検索対象となる文書ファイル中に存在する属性名について、属性名とその属性名を表現する自然言語表現との対の集合を格納しておく属性名格納辞書と、

検索対象となる文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組の集合を格納しておく属性値格納辞書と、

利用者が、自然言語で表現した検索要求文を入力すると、前記入力文を先頭から順に探査し、属性名格納辞書を参照して、属性名を表現する自然言語表現が含まれていたら、その自然言語表現と対である属性名を属性名インデックスとして出力し、属性値格納辞書を参照して、属性値を表現する自然言語表現が含まれていたら、その自然言語表現と３つ組である属性値と属性名との対の集合を属性値インデックスとして出力することを順次行うキーワード抽出部と、

キーワード抽出部の出力を入力し、先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣通しに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部と、

文書ファイルの内容と属性値インデックスとを入力すると、前記文書ファイルの内容中に、前記属性値インデックス中の属性名を含むタグが存在するかどうか調べ、存在する場合は、そのタグと対で存在する属性値を取り出し、その値が前記属性値インデックス中の属性値と等しいかどうか調べ、等しい場合は、合格の出力をし、そうでない場合は不合格の出力をする文書内容検査部と、

文書ファイルの内容と１つ以上の属性値インデックスとを入力すると、前記属性値インデックスから１つずつ取り出し、前記文書ファイルの内容と前記取り出した属性値インデックスとを１つずつ文書内容検査部に渡していき、すべての属性値インデックスに対してその出力が合格のときは、合格を出力し、そうでないときは不合格を出力する統合文書内容検査部と、

前記文書ファイル名辞書を参照して、１つずつ文書ファイルの内容を取り出し、前記文書の内容とキーワードフィルタ部の出力のうちの属性値インデックスの部分とを統合文書内容検査部に渡し、前記統合文書内容検査部の出力を受け取ることを前記１つずつ取り出した文書ファイルのすべてに対して行い、前記出力が合格の文書ファイルの名前のみを出力する合格文書ファイル名選別部と、

文書ファイル名と前記文書ファイル名の内容とキーワードフィルタ部の出力である属性名インデックスとを入力すると、前記属性名インデックスのうちの１つを取り出し、与えられた前記文書ファイルの内容中に、前記取り出した属性名を含むタグが存在するかどうか調べ、存在する場合は、その属性名のタグの値と前記入力した文書ファイル名とを利用者に表示し、存在しない場合には何も出力しないことを、前記入力した属性名インデックスのそれぞれに対して行う文書内容出力部と、

前記合格文書ファイル名選別部の出力である文書ファイル名の集合を入力し、文書ファイル名格納辞書を参照して、前記入力した文書ファイル名の集合の要素を１つずつ取り出し、文書内容出力部に渡すことを、前記入力中の文書ファイル名のすべてに対して行うことを繰り返す文書内容出力制御部とを備えることを特徴とする文書ファイル検索装置。

【請求項５】　属性の属性名とその属性の属性値との対を内蔵する文書ファイルを検索対象文書ファイルとし、検索対象文書ファイルから、利用者が自然言語で指定した検索条件に適合する部分を検索する文書ファイル検索装置を構成するコンピュータを、

自然言語で表現した検索要求文を先頭から順に探査し、属性名を表現する自然言語表現に対してはその属性名を属性名インデックスとして出力し、属性値を表現する自然言語表現に対してはその属性値と属性名との対を属性値インデックスとして出力することを順次行うキーワード抽出部、

前記キーワード抽出部の出力を入力して先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部、

検索対象文書ファイル中に、前記キーワードフィルタ部から出力された全ての属性値インデックスの属性名と属性値との対が内蔵されているか否かを調べ、内蔵されている場合、前記キーワードフィルタ部から出力された属性名インデックスの属性名に対応する属性値を検索対象文書ファイルから検索して出力する検索手段、

として機能させるプログラムを記録した機械読み取り可能な記録媒体。

【請求項６】　文書中に書かれた意味を表現する属性名のついたタグとその属性の属性値との対を内蔵する文書ファイルを検索対象文書ファイルとし、検索対象文書ファイルから、利用者が自然言語で指定した検索条件に適合する部分を検索する文書ファイル検索装置を構成するコンピュータを、

検索対象文書ファイル中に存在する属性名について、属性名とその属性名を表現する自然言語表現との対の集合を格納しておく属性名格納辞書、

検索対象文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組の集合を格納しておく属性値格納辞書、

自然言語で表現した検索要求文を先頭から順に探査し、属性名格納辞書を参照して、属性名を表現する自然言語表現が含まれていたら、その自然言語表現と対である属性名を属性名インデックスとして出力し、属性値格納辞書を参照して、属性値を表現する自然言語表現が含まれていたら、その自然言語表現と３つ組である属性値と属性名との対を属性値インデックスとして出力するキーワード抽出部、

キーワード抽出部の出力を入力し、先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部、

検索対象文書ファイル中に、前記キーワードフィルタ部から出力された全ての属性値インデックスの属性名と属性値との対に対応するタグの対が内蔵されているか否かを調べ、内蔵されている場合、前記キーワードフィルタ部から出力された属性名インデックスの属性名を持つタグの属性値を検索対象文書ファイルから検索して出力する検索手段、

として機能させるプログラムを記録した機械読み取り可能な記録媒体。

【請求項７】　文書中に書かれた意味を表現する属性名のついたタグとその属性の値との対を複数個内蔵する文書ファイルの集合から、利用者が自然言語で指定した検

索条件を満足する文書ファイルを選択してその適合する部分を表示する文書ファイル検索装置を構成するコンピュータを、
検索対象となるすべての文書ファイルの名前と存在位置とを格納する文書ファイル名辞書、
検索対象となる文書ファイル中に存在する属性名について、属性名とその属性名を表現する自然言語表現との対の集合を格納しておく属性名格納辞書、
検索対象となる文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組の集合を格納しておく属性値格納辞書、
利用者が、自然言語で表現した検索要求文を入力すると、前記入力文を先頭から順に探査し、属性名格納辞書を参照して、属性名を表現する自然言語表現が含まれていたら、その自然言語表現と対である属性名を属性名インデックスとして出力し、属性値格納辞書を参照して、属性値を表現する自然言語表現が含まれていたら、その自然言語表現と３つ組である属性値と属性名との対の集合を属性値インデックスとして出力することを順次行うキーワード抽出部、
キーワード抽出部の出力を入力し、先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣通しに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部、
文書ファイルの内容と属性値インデックスとを入力すると、前記文書ファイルの内容中に、前記属性値インデックス中の属性名を含むタグが存在するかどうか調べ、存在する場合は、そのタグと対で存在する属性値を取り出し、その値が前記属性値インデックス中の属性値と等しいかどうか調べ、等しい場合は、合格の出力をし、そうでない場合は不合格の出力をする文書内容検査部、
文書ファイルの内容と１つ以上の属性値インデックスを入力すると、前記属性値インデックスから１つずつ取り出し、前記文書ファイルの内容と前記取り出した属性値インデックスとを１つずつ文書内容検査部に渡していき、すべての属性値インデックスに対してその出力が合格のときは、合格を出力し、そうでないときは不合格を出力する統合文書内容検査部、
前記文書ファイル名辞書を参照して、１つずつ文書ファイルの内容を取り出し、前記文書の内容とキーワードフィルタ部の出力のうちの属性値インデックスの部分とを統合文書内容検査部に渡し、前記統合文書内容検査部の出力を受け取ることを前記１つずつ取り出した文書ファイルのすべてに対して行い、前記出力が合格の文書ファイルの名前のみを出力する合格文書ファイル名選別部、
文書ファイル名と前記文書ファイル名の内容とキーワードフィルタ部の出力である属性名インデックスとを入力すると、前記属性名インデックスのうちの１つを取り出し、与えられた前記文書ファイルの内容中に、前記取り出した属性名を含むタグが存在するかどうか調べ、存在する場合は、その属性名のタグの値と前記入力した文書ファイル名とを利用者に表示し、存在しない場合には何も出力しないことを、前記入力した属性名インデックスのそれぞれに対して行う文書内容出力部、
前記合格文書ファイル名選別部の出力である文書ファイル名の集合を入力し、文書ファイル名格納辞書を参照して、前記入力した文書ファイル名の集合の要素を１つずつ取り出し、文書内容出力部に渡すことを、前記入力中の文書ファイル名のすべてに対して行うことを繰り返す文書内容出力制御部、
として機能させるプログラムを記録した機械読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は文書ファイル検索装置に関し、特に自然言語による検索問い合わせを可能とした文書ファイル検索装置に関する。

【０００２】

【従来の技術】一般に情報検索において利用者の検索意図をより精密に表現させようとする場合には、日本語や英語のような自然言語によってそれを表現させる方法が有効である。データベースに対する検索を自然言語で行うシステムは既に存在し、自然言語インタフェースと呼ばれている（参考文献：ディベロッピングアナチュラルランゲージインタフェースツーコンプレックスデータ、ジー．ジー．ヘンドリックス他，“Developing a Natural Language Interface to Complex Data”，ACM Trans. on Database Systems, 1978.）。

【０００３】従来の自然言語インタフェースは、利用者の自然言語による検索問い合わせを解釈して、その問い合わせをデータベースの検索言語（SQL）の検索式に変換し、その検索式をデータベースシステムに送り、データベースシステムから戻された検索結果を利用者に提示するものである。

【０００４】しかし、従来の自然言語インタフェースは、既に商用化が始まって20年以上たったのにも関わらず、まだ実用のレベルに達していない。その理由の１つは、自然言語インタフェースシステムが利用者の自由な問い合わせを解釈することができず、そのシステムが許容する構文や語彙が明確に限定されているので、結局は利用者はどういう言い回しが使えるかを覚えなくてはならないためである。従って、自然言語インタフェースといっても複雑なコマンド体系と変わらない。つまり、従来の自然言語インタフェースは、利用者の自由な言い回しを受け付けることが出来ないと言うことが問題であった（参考文献：ディベロッピングアナチュラルランゲージインタフェースツーコンプレックスデータ、ジー．ジー．ヘンドリックス他，“Developing a Natural Langu

age Interface to Complex Data", ACM Trans. on Database Systems, 1978.)。とりわけ、簡単な質問については、利用者は完璧な自然言語でなく、簡潔な言い回しを使いたいのに、それが許されないということが問題であった。簡潔な言い回しの例としては、キーワード列による表現や非文法的な言い回し、あるいは自然言語文の一部、等がある。

【０００５】そこで本出願人は、先の特許出願において、簡潔な言い回しの自然言語による検索を実現する「自然言語解釈方法」を提案した（特開平５－６７１３６号公報）。これは、属性の属性名とその属性の属性値との組の集まりであるデータベースのテーブルを検索対象とし、自然言語による問い合わせ文中の各単語を属性名と属性値とその他とに分類し、属性名に分類された単語は応答属性名として保存し、属性値に分類された単語はその属性値とそれに対応する属性名とを組にして条件属性値組群として保存し、この保存された条件属性値組群中の属性値と属性名との組が全てテーブル中に存在する場合、前記応答属性名として保存された各属性名に対応する前記テーブル中の属性値を、問い合わせに対する回答として出力するものである。なお、これに類似する従来技術として、やはり本出願人によって先に出願された特開平５－２４２１４７号公報にかかる「自然言語解釈方法」がある。

【０００６】他方、最近におけるWorld Wide Web (WWW) の利用の拡大に伴い、WWW 上での検索技術の重要性が高まっている。WWW の利用者がWWW 上で情報検索をするときに使う典型的なツールはサーチエンジンである。この例としては、Altavista, Infoseek, Lycosなどが有名である。しかし、サーチエンジンでは、キーワードを組み合わせる検索の形式なので、利用者の検索の意図が直接的に反映させられないことが多い。例えば、ホテルに関する情報を探すときに、値段が150 ドルでペットを連れて行くのが可能で、しかもサンフランシスコ近辺にあるホテルのホームページを見つけたいときに、そのような検索意図をキーワードの並びのみで表現することは不可能である。仮に、(1) 式で表現するように

150 ドル、サンフランシスコ、ホテル、ペット可能　　… (1)

とキーワードを並べたら、非常に大量のホームページのリストが出力される。

【０００７】自然言語インタフェースシステムを導入しWWW のホームページを検索できれば、上の例のような精密な検索条件を素直に表現し、適切なホームページのみが検索できることになる。しかしながら、WWW 上のホームページの検索に対し自然言語インタフェースを適用した例は見当たらない。

【０００８】なお、WWW のホームページに対する検索技術の他の例として、特開平10-40262号公報に記載された「情報検索装置」があるが、これは、感性表現データをキーワードにした検索を可能にすることで、明確な検索対象または検索条件を持たない利用者の感性に合った情報検索を目的としているため、上の例のような精密な検索条件による検索には向いていない。

【０００９】

【発明が解決しようとする課題】上述した特開平５－６７１３６号公報に記載された技術は、単純な方法で自然言語による問い合わせを解釈することができるので、自然言語インタフェースシステムを実用化する上で有効な手段となり得るが、未だ解決すべき課題が残されている。それは、自然言語による問い合わせ文中に或る属性名が存在する場合、それが必ず応答属性名として扱われ、問い合わせに対する回答中に含められるため、回答が冗長になる場合があることである。

【００１０】例えば、「属性名＝書名，その属性値＝人間失格、属性名＝著者，その属性値＝太宰治」を持つテーブルに対して、「書名＝人間失格の著者は？」という問い合わせを行った場合、まず、「書名」が属性名と判定されて応答属性名として保存され、次いで「人間失格」が属性値と判定されて属性値「人間失格」とその属性名である「書名」との組が条件属性値組群として保存され、次いで、「著者」が属性名と判定されて応答属性名として保存される。そして、条件属性値組群中の属性値「人間失格」と属性名「書名」との組を有する前記テーブルが検索され、そのテーブルから応答属性名「書名」と「著者」とに対応する属性値「人間失格」と「太宰治」とが検索されて出力される。つまり、「人間失格」をも出力している分、回答が冗長になっている。

【００１１】そこで本発明の目的は、自然言語による検索問い合わせに対する回答の冗長性を極力無くすことにある。

【００１２】また、本発明の他の目的は、WWW のホームページに対しても自然言語による検索問い合わせを可能にすることにある。

【００１３】

【課題を解決するための手段】（１）第１の発明

上述した特開平５－６７１３６号公報に記載された技術において、回答が冗長になっている理由は、自然言語による問い合わせ文中に属性名が存在する場合、それに対応する属性値を利用者が問い合わせ文中で記述しているにもかかわらず、一律に応答属性名として扱っているためである。そこで、本発明では、自然言語による問い合わせ文中で検索条件を指定するために或る属性名とそれに対応する属性値とを記述する場合、利用者はそれらを互いに隣接して記述する傾向にある点に着目し、同一の属性の属性名と属性値とが隣どうしに現れる場合にその属性名を応答属性名に含めないようにしている。より具体的には、属性の属性名とその属性の属性値との対を内蔵する文書ファイルを検索対象文書ファイルとし、検索

対象文書ファイルから、利用者が自然言語で指定した検索条件に適合する部分を検索する文書ファイル検索装置において、自然言語で表現した検索要求文を先頭から順に探査し、属性名を表現する自然言語表現に対してはその属性名を属性名インデックスとして出力し、属性値を表現する自然言語表現に対してはその属性値と属性名との対を属性値インデックスとして出力することを順次行うキーワード抽出部と、前記キーワード抽出部の出力を入力して先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部と、検索対象文書ファイル中に、前記キーワードフィルタ部から出力された全ての属性値インデックスの属性名と属性値との対が内蔵されているか否かを調べ、内蔵されている場合、前記キーワードフィルタ部から出力された属性名インデックスの属性名に対応する属性値を検索対象文書ファイルから検索して出力する検索手段とを備えている。

【００１４】更に、検索要求文中のどの自然言語表現が属性名を表現し、またどの自然言語表現が属性値を表現しているかを正確に判定できるようにするために、検索対象文書ファイル中に存在する属性名について、属性名とその属性名を表現する自然言語表現との対を格納しておく属性名格納辞書と、検索対象文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組を格納しておく属性値格納辞書とを備え、前記キーワード抽出部は、自然言語で表現した検索要求文を先頭から順に探査し、属性名格納辞書を参照して、属性名を表現する自然言語表現が含まれていたら、その自然言語表現と対である属性名を属性名インデックスとして出力し、属性値格納辞書を参照して、属性値を表現する自然言語表現が含まれていたら、その自然言語表現と３つ組である属性値と属性名との対の集合を属性値インデックスとして出力する構成を有する。

*【００１５】このように構成された本発明の文書ファイル検索装置にあっては、利用者が自然言語で表現した検索要求文を入力すると、まずキーワード抽出部が、検索要求文を先頭から順に探査し、属性名を表現する自然言語表現に対してはその属性名を属性名インデックスとして出力し、属性値を表現する自然言語表現に対してはその属性値と属性名との対を属性値インデックスとして出力し、次いでキーワードフィルタ部が、キーワード抽出部の出力を入力して先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、属性名インデックスを削除し、次いで、検索手段が、検索対象文書ファイル中に、キーワードフィルタ部から出力された全ての属性値インデックスの属性名と属性値との対が内蔵されているか否かを調べ、内蔵されている場合、キーワードフィルタ部から出力された属性名インデックスの属性名に対応する属性値を検索対象文書ファイルから検索して出力することにより、利用者への回答が冗長になるのを防いでいる。

【００１６】（２）第２の発明

WWW 上のホームページの検索に対し自然言語インタフェースを適用するのが困難であった理由は、WWW のホームページの中身が自然言語で書かれた文章や図から構成されるファイルであり、データベースのように、属性名と属性値の集合でないことである。つまり、従来の自然言語インタフェースシステムが対象とするデータベースは、属性名と属性値の集合だったため、従来の技術の項の例で出てくるホテルのデータベースがあるとすると、

（名前：Ｘホテル、値段：＄150 、ペット：可能、場所：サンフランシスコ）

（名前：Ｙホテル、値段：＄200 、ペット：不可、場所：ロサンゼルス）

（名前：Ｚホテル、値段：＄180 、ペット：不可、場所：シアトル）

のような形態で格納されており、利用者の問い合わせは、(2) 式で表現するように

「値段＝$150　かつ　ペット＝可能　かつ　場所＝サンフランシスコ」　… (2)

のように変換できた。これは、SQL 言語にそのまま変換される。しかし、WWW のホームページは、通常このような属性名と属性値の情報が入っていないので、SQL の式で表現できるような検索式に変換できない。つまり、従来のWWW のホームページ作成言語はHTMLである（参考文献：ワールドワイドウェブコンソーシアムのホームページ、URL http://www.w3.org)。HTMLでは、文書ファイル中に、その構成を表現するための属性名と属性値の組が内蔵されている。例えば、図６に示すのが、HTMLファイルの例である。ここで、＜と＞に囲まれたものが属性タグであり、単独で使われるもの（例：＜Ｂ＞）と、開始タグ（例：<TR>）と終了タグ（例：</TR >）の対で使われるものとがある。HTMLのタグの特徴は、それが文書ファイル中の外見の表現を定義するのに限定されていることである。例えば、表的な表現にするタグは<TABLE >であり、改行を表すタグは＜Ｐ＞で表現される。このようなHTMLファイルをWWW ブラウザに読み込ませると、図７に示すような形態になってユーザに出力表示される。しかし、HTMLでは、文書中の意味を表現する為のタグを定義することは出来ない。

【００１７】そこで本発明では、WWW の文書ファイルに、その文書ファイル中の意味を表現する属性名と属性値との組を内蔵させる。具体的には、例えば、ファイル中に文書の内容を属性タグとその属性値の対の集合で表

現できるようにXML (Extensible Markup Language)で文書を記述する（参考文献：ワールドワイドウェブコンソーシアムのホームページ、「エクステンシブル　マークアップ　ランゲージ 1.0」http://www.w3.org/TR/PR-xml-971208）。XML は、WWW の標準を決める機関であるワールドワイドウェブコンソーシアム（参考文献：ワールドワイドウェブコンソーシアムのホームページ、URL http://www.w3.org)によって1997年12月にその仕様が提案された。XML で記述された文書では、文書の内容を機械が可読になって内容による検索が可能になる。そこで、本発明ではそのことを利用してWWW のホームページに対して自然言語による検索問い合わせを実現する。

【００１８】具体的には、文書中に書かれた意味を表現する属性名のついたタグとその属性の属性値との対を内蔵する文書ファイルを検索対象文書ファイルとし、検索対象文書ファイルから、利用者が自然言語で指定した検索条件に適合する部分を検索する文書ファイル検索装置において、検索対象文書ファイル中に存在する属性名について、属性名とその属性名を表現する自然言語表現との対の集合を格納しておく属性名格納辞書と、検索対象文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組の集合を格納しておく属性値格納辞書と、自然言語で表現した検索要求文を先頭から順に探査し、属性名格納辞書を参照して、属性名を表現する自然言語表現が含まれていたら、その自然言語表現と対である属性名を属性名インデックスとして出力し、属性値格納辞書を参照して、属性値を表現する自然言語表現が含まれていたら、その自然言語表現と３つ組である属性値と属性名との対の集合を属性値インデックスとして出力するキーワード抽出部と、キーワード抽出部の出力を入力し、先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部と、検索対象文書ファイル中に、前記キーワードフィルタ部から出力された全ての属性値インデックスの属性名と属性値との対に対応するタグが内蔵されているか否かを調べ、内蔵されている場合、前記キーワードフィルタ部から出力された属性名インデックスの属性名を持つタグの属性値を検索対象文書ファイルから検索して出力する検索手段とを備えている。

【００１９】このように構成された本発明の文書ファイル検索装置にあっては、利用者が自然言語で指定した検索要求文を入力すると、キーワード抽出部が、検索要求文を先頭から順に探査し、属性名を表現する自然言語表現が含まれていたら、その属性名を属性名インデックスとして出力し、属性値を表現する自然言語表現が含まれていたら、その属性値と属性名との対の集合を属性値インデックスとして出力し、次いで、キーワードフィルタ部が、キーワード抽出部の出力を入力し、先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、前記属性名インデックスを削除し、次いで、検索手段が、検索対象文書ファイル中に、キーワードフィルタ部から出力された全ての属性値インデックスの属性名と属性値との対に対応するタグが内蔵されているか否かを調べ、内蔵されている場合、キーワードフィルタ部から出力された属性名インデックスの属性名を持つタグの属性値を検索対象文書ファイルから検索して出力する。

【００２０】また、予め登録された多数の文書ファイルの内から利用者が自然言語で入力した検索条件を満たす文書ファイルのみを選別し、さらにその中の利用者が必要な部分を利用者に表示できるようにするために、文書中に書かれた意味を表現する属性名のついたタグとその属性の値との対を複数個内蔵する文書ファイルの集合から、利用者が自然言語で指定した検索条件を満足する文書ファイルを選択してその適合する部分を表示する文書ファイル検索装置において、検索対象となるすべての文書ファイルの名前と存在位置とを格納する文書ファイル名辞書と、検索対象となる文書ファイル中に存在する属性名について、属性名とその属性名を表現する自然言語表現との対の集合を格納しておく属性名格納辞書と、検索対象となる文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組の集合を格納しておく属性値格納辞書と、利用者が、自然言語で表現した検索要求文を入力すると、前記入力文を先頭から順に探査し、属性名格納辞書を参照して、属性名を表現する自然言語表現が含まれていたら、その自然言語表現と対である属性名を属性名インデックスとして出力し、属性値格納辞書を参照して、属性値を表現する自然言語表現が含まれていたら、その自然言語表現と３つ組である属性値と属性名との対の集合を属性値インデックスとして出力することを順次行うキーワード抽出部と、キーワード抽出部の出力を入力し、先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合のみ、前記属性名インデックスを削除し、それ以外の部分はそのまま出力するキーワードフィルタ部と、文書ファイルの内容と属性値インデックスとを入力すると、前記文書ファイルの内容中に、前記属性値インデックス中の属性名を含むタグが存在するかどうか調べ、存在する場合は、そのタグと対で存在する属性値を取り出し、その値が前記属性値インデックス中の属性値と等しいかどうか調べ、等しい場合は、合格の出力をし、そうでない場合は不合格の出力をする文書内容検査部と、文書ファイルの内容と１つ以上の属性値インデックスとを入力すると、前記属性値インデックスから１つずつ取り出し、前記文書ファイルの内容と前記取り出

した属性値インデックスを１つずつ文書内容検査部に渡していき、すべての属性値インデックスに対してその出力が合格のときは、合格を出力し、そうでないときは不合格を出力する統合文書内容検査部と、文書ファイル名辞書を参照して、１つずつ文書ファイルの内容を取り出し、前記文書の内容とキーワードフィルタ部の出力のうちの属性値インデックスの部分とを統合文書内容検査部に渡し、前記統合文書内容検査部の出力を受け取ることを前記１つずつ取り出した文書ファイルのすべてに対して行い、前記出力が合格の文書ファイルの名前のみを出力する合格文書ファイル名選別部と、文書ファイル名と前記文書ファイル名の内容とキーワードフィルタ部の出力である属性名インデックスとを入力すると、前記属性名インデックスのうちの１つを取り出し、与えられた前記文書ファイルの内容中に、前記取り出した属性名を含むタグが存在するかどうか調べ、存在する場合は、その属性名のタグの値と前記入力した文書ファイル名とを利用者に表示し、存在しない場合には何も出力しないことを、前記入力した属性名インデックスのそれぞれに対して行う文書内容出力部と、前記合格文書ファイル名選別部の出力である文書ファイル名の集合を入力し、文書ファイル名格納辞書を参照して、前記入力した文書ファイ*

*ル名の集合の要素を１つずつ取り出し、文書内容出力部に渡すことを、前記入力中の文書ファイル名のすべてに対して行うことを繰り返す文書内容出力制御部とを備えている。

【００２１】このように構成された本発明の文書ファイル検索装置の作用を、その理解を容易にするために、例を使って説明する。まず、利用者が検索する対象となるWWW文書ファイルとして、図５（ａ），（ｂ）に示したものを使用する。図５の文書ファイル中には、文章テキストの他に、属性の属性名とその属性の属性値との対が含まれている。また、利用者の検索文の例として、次の文を使う。

検索入力文：「値段が150 ドルでペットを連れて行くのが可能で、しかも、サンフランシスコ近辺にあるホテルの情報を見つけたい」

【００２２】まず、第１段階では、入力文をキーワード列に変換する。キーワードの種類としては、２種類存在する。１つ目は、属性名を参照する自然言語表現であり、属性名インデックスと呼ぶ。２つ目は、属性値を参照する自然言語表現であり、属性値インデックスと呼ぶ。

【００２３】

生成されたキーワード列：

｜

価格（「価格」の属性名インデックス）、
$150ドル（「価格」の属性値インデックス）、
ペット（「ペット可能性」の属性名インデックス）、
可能（「ペット可能性」の属性値インデックス）、
サンフランシスコ（「場所」の属性値インデックス）、
ホテル（「ホテル名」の属性名インデックス）

｜　　…(4)

【００２４】次に、属性名インデックスと属性値インデックスの並び順を参照して、冗長な部分の統合を行う。同一の属性に対する属性名インデックスと属性値インデ※

※ックスとが隣りどうしに並んでいるときには、属性名インデックスの方を削除する。上のキーワード列例は、次のように圧縮される。

圧縮されたキーワード列：

｜

$150ドル（「価格」の属性値インデックス）、
可能（「ペット可能性」の属性値インデックス）、
サンフランシスコ（「場所」の属性値インデックス）、
ホテル（「ホテル名」の属性名インデックス）

｜　　…(5)

【００２５】次に、抽出したキーワード列を解釈する。属性値インデックスは、それが参照する属性の値として、属性値インデックスが保持する値を取ること、という条件式と解釈する。例えば、

$150 ドル（「価格」の属性値インデックス）

は、

★「「価格」属性の値　＝ $150 」

という解釈をする。

【００２６】複数の属性値インデックスが存在する場合は、それらの解釈を論理積したものが全体の条件式となる。上の例では、以下のようになる。

★

全体の条件式

｜

「「価格」属性の値 = $150 」　かつ

15　　16

「「ペット可能性」属性の値 = 可能」　かつ
「「場所」属性の値 = サンフランシスコ」
|　　…(6)

【００２７】属性名インデックスは、それが参照する属性の値を出力せよ、という解釈になる。上の例では、以＊下のようになる。＊

検索部分の特定
|
ホテル（「ホテル名」の属性名インデックス）」
|　　…(7)

【００２８】この意味は、「「ホテル名」属性の値を出力せよ」という解釈となる。複数の属性名インデックスがあるときは、それら複数の属性名インデックスを順次出力せよ、という意味になる。

【００２９】入力文全体の解釈は、属性値インデックスから生成される検索条件式を満足するWWW 上の文書ファイルを選択し、次に、それらの文書ファイル中から属性名インデックスの解釈で指定される属性名の値を抽出してそれを利用者に表示すれば良い。

【００３０】

【発明の実施の形態】図１を参照すると、本発明の実施の形態の文書ファイル検索装置１００は、文書ファイル名辞書１と、属性名格納辞書２と、属性値格納辞書３と、キーワード抽出部４と、キーワードフィルタ部５と、文書内容検査部６と、統合文書内容検査部７と、合格文書ファイル名選別部８と、文書内容出力部９と、文書内容出力制御部10とから構成され、キーボード等の入力装置１０１、ＣＲＴディスプレイ等の表示装置１０２およびインターネット１０３に接続されている。

【００３１】文書ファイル名辞書１には、検索対象となるすべての文書ファイルの名前とその物理的な位置とが格納されている。検索対象となる文書ファイルがHTMLやXMLで記述されている場合には、文書ファイルは、世界中のWWW サーバに分散していることも可能である。その場合、文書ファイルの位置は、「http://.....」というURL 記述になる。

【００３２】属性名格納辞書２には、検索対象となる文書ファイル中に存在する属性タグの属性名とその属性名を表現する自然言語表現との対が登録されている。ある属性名を参照する自然言語表現の中の最も基本的なものは、その属性名そのものである。例えば、「ホテル」という属性名を参照する自然言語表現としては、「ホテル」である。しかし、それ以外にも、「ホテル」を参照する表現がある。例えば、「宿泊場所」、「泊まるところ」などの表現がある。これらが、下記の表１でで示すような対になって登録される。

【００３３】

【表１】

属性名	自然言語表現
ホテル	ホテル
ホテル	宿泊場所
ホテル	泊まるところ

【００３４】属性値格納辞書３には、検索対象となる文書ファイル中に存在する属性値について、属性値とその属性値に対応する属性名とその属性値を表現する自然言語表現との３つ組が格納される。ある属性値を参照する自然言語表現としてもっとも基本的なものは、その属性値そのものである。例えば、「Ｘホテル」という属性値を参照する自然言語表現としては、「Ｘホテル」そのものがありこれ以外にはないかもしれない。しかし、別の例では、「ペット」属性の属性値を表わす自然言語表現としては、「可能」の他に「動物連れ込みOK」「ペット同伴OK」「犬猫可」のような表現も登録しておいてもよい。属性値格納辞書３には、下記の表２で示すように３つ組でデータが格納される。

【００３５】

【表２】

属性値	自然言語表現	対応する属性名
可能	可能	ペット
可能	動物連れ込みOK	ペット
可能	ペット同伴OK	ペット
可能	犬猫可	ペット
不可	動物不可	ペット

【００３６】キーワード抽出部４は、自然言語表現による入力条件検索文を入力装置101 を通じて利用者から受け取ると、属性名格納辞書２と属性値格納辞書３とを参照して、その中の自然言語表現として登録されている表現が入力条件検索文中にないかどうかを調べる。あった場合には、それが属性名の場合には、属性名のみを出力する。この出力のことを属性名インデックスと呼ぶ。他方、それが属性値の場合には、属性値と対応する属性名との対を出力する。この出力のことを属性値インデックスと呼ぶ。これらは、入力条件検索文の先頭から調べていき、マッチするものが見つかったら、その順番に出力していく。

【００３７】キーワードフィルタ部５は、キーワード抽出部４の出力をそのまま受け取り、先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合は、その属性名インデックスを削除し、それ以外の部分はそのまま素通しで出力する。

【００３８】文書内容検査部６は、統合文書内容検査部７から文書ファイルの内容である文字列と属性値インデックスとを入力として受け付ける。入力として受け取った文書ファイルの内容文字列中に、受け取った属性値インデックス中の属性名を含むタグが存在するかどうか調べ、存在する場合は、そのタグと対で存在する属性値を取り出し、その値がこの属性値インデックス中の属性値と等しいかどうか調べ等しい場合は、合格の出力をし、そうでない場合は不合格の出力をする。文書内容検査部６は、統合文書内容検査部７から呼び出されて動作する一種のサブルーチン的役割を果たしている。

【００３９】統合文書内容検査部７は、合格文書ファイル名選別部８から文書ファイルの内容である文字列と１つ以上の属性値インデックスとを入力として受け付ける。与えられた属性値インデックスは、１つ１つが「属性値インデックス中に記述された属性の値として、属性値インデックス中に記述された値をとらねばならない」という条件式を表現していると見做す。統合文書内容検査部７の役割は、与えられた文字列中から、属性値インデックスに記述された属性表現を見つけて、その条件が満足されているかを調べることである。入力として与えられた１つ以上の属性値インデックスのすべての条件を満足すれば、「合格」という値を出力し、そうでない場合は、「不合格」という値を出力する。実際に、文書ファイルの内容である文字列が１つの属性値インデックスの条件を満足するかどうかを判定するのは、文書内容検査部６が行う。統合文書内容検査部７は、複数の属性値インデックスがあった場合に、属性値インデックス１つずつを文書内容検査部６に順々に渡していく一種のループ制御を行っている。統合文書内容検査部７も、合格文書ファイル名選別部８から呼び出されるサブルーチン的役割である。

【００４０】合格文書ファイル名選別部８は、文書ファイル名辞書１を参照して、必要に応じてインターネット１０３を通じて世界中に分散しているWWW サーバをアクセスして１つずつ文書ファイルの内容を取り出し、この文書の内容とキーワードフィルタ部５の出力のうち属性値インデックスの部分とを統合文書内容検査部７に渡し、統合文書内容検査部７の出力を受け取る。ここで出力としては、「合格」または「不合格」の値が返される。この処理を文書ファイル名辞書１に登録されているすべてのファイルに対して行い、統合文書内容検査部７の出力が「合格」だったファイルに対してのみ、文書ファイル名を文書内容出力制御部10に出力する。

【００４１】文書内容出力部９は、文書ファイル名とこのファイルの内容とキーワードフィルタ部５の出力である１つ以上の属性名インデックスとを入力する。入力した属性名インデックスのうちの１つを取り出し、入力した文書ファイルの内容中に、この属性名インデックス中の属性名を含むタグが存在するかどうか調べ、存在する場合は、その属性名タグに対応する属性値タグの値と入力した文書ファイル名との対を表示装置102 を通じて利用者に表示し、存在しなかった場合には何も出力しないという処理を、入力したすべての属性名インデックスのそれぞれに対して行う。文書内容出力部９は、文書内容出力制御部10によってサブルーチン的に呼び出される役

19

割をしている。なお、属性値タグの値と文書ファイル名との対を出力する代わりに、属性値タグの値と文書ファイルの位置情報とを表示するようにしても良く、また、属性値タグの値と文書ファイル名とその位置情報とを表示するようにしても良い。

【００４２】文書内容出力制御部10は、合格文書ファイル名選別部８の出力である文書ファイル名の集合をそのまま自身の入力とし、文書ファイル名辞書１を参照して、入力した文書ファイル名の集合中の文書ファイルの内容を必要に応じてインターネット103 を通じてWWW サーバをアクセスして１つずつ取り出し、文書ファイル名およびキーワードフィルタ部５で生成された属性名インデックスとともに文書内容出力部９に渡すことを、入力中の文書ファイル名のすべてに対して行うことを繰り返すものである。つまり、入力として合格した文書ファイル名を３つ受け取った場合には、３回文書内容出力部９を呼び出すことになる。なお、合格文書ファイル名選別部８がインターネット103 を通じてWWWWサーバから取り込んだ文書ファイルの内容が磁気ディスク装置等に保存されている場合、文書内容出力制御部10はその内容を利用することで、インターネット103 へのアクセス回数を減らすことができる。

【００４３】図２および図３は文書ファイル検索装置100 の処理例を示すフローチャートである。以下、本実施の形態の動作について説明する。

【００４４】キーワード抽出部４は、入力装置101 を通じて利用者から自然言語表現による検索入力文を受け付けると（ステップＳ１）、属性名格納辞書２と属性値格納辞書３とを参照して、その中の自然言語表現として登録されている表現が検索入力文にないかどうかを、検索入力文の先頭から順に調べ、あった場合には、それが属性名のときは属性名のみを含む属性名インデックスを出力し、それが属性値のときは属性値と対応する属性名との対を含む属性値インデックスを出力する（ステップＳ２）。

【００４５】次にキーワードフィルタ部５は、キーワード抽出部４から出力されたインデックスの並びを検査し、同一の属性の属性名インデックスと属性値インデックスとが連続している箇所を検出し、その箇所の属性名インデックスを削除する（ステップＳ３）。

【００４６】次に合格文書ファイル名選別部８は、文書ファイル名辞書１中の１つの文書ファイル名に注目し、その文書ファイル名の文書の内容を取り出して、キーワードフィルタ部５から出力された全ての属性値インデックスとともに統合文書内容検査部７に渡し、合否を判定させる（ステップＳ４）。

【００４７】統合文書内容検査部７は、渡された文書内容を検査するために、まず渡された属性値インデックスの１つに注目し、この属性値インデックスと文書ファイルの内容とを文書内容検査部６に渡し、合否を判定させる（ステップＳ５）。

20

【００４８】文書内容検査部６は、渡された文書ファイルの内容中に、渡された属性値インデックスに含まれる属性名を持つ属性名タグが存在し、かつ、その存在した属性名タグと対になっている属性値タグの値が、渡された属性値インデックスに含まれる属性値と一致するかを検査し、一致する場合には合格を、そのような属性名タグが存在しないか或いは存在してもその属性値が一致しない場合には不合格を、統合文書内容検査部７に通知する（ステップＳ６）。

【００４９】統合文書内容検査部７は、文書内容検査部６から合格が通知された場合（ステップＳ７でＹＥＳ）、合格文書ファイル名選別部８から通知された全ての属性値インデックスについて検査し終えたか否かを調べ、未だ検査し終えていないときは（ステップＳ８でＮＯ）、残りの属性値インデックスの１つに注目を移し、その属性値インデックスと文書ファイルの内容とを文書内容検査部６に渡し、合否を判定させる（ステップＳ９）。そして、全ての属性値インデックスについて文書内容検査部７で合格の判定が出た場合（ステップＳ８でＹＥＳ）、合格文書ファイル名選別部８に合格を通知し、合格文書ファイル名選別部８は当該文書ファイルを合格文書ファイルとし（ステップＳ10）、ステップＳ11へと進む。他方、文書内容検査部６から不合格が通知された場合（ステップＳ７でＮＯ）、統合文書内容検査部７は合格文書ファイル名選別部８に不合格を通知し、合格文書ファイル名選別部８はステップＳ11へと進む。

【００５０】合格文書ファイル名選別部８は、１つの文書ファイルについての合否判定が終わると、文書ファイル名辞書１中に未処理の文書ファイルが残っている場合（ステップＳ11でＹＥＳ）、その内の１つの文書ファイル名に注目を移し（ステップＳ12）、先の文書ファイルと同様にして合否の判定を下す。

【００５１】文書ファイル名辞書１中の全ての文書ファイルに対する合否判定を終えると（ステップＳ11でＹＥＳ）、合格文書ファイル名選別部８は、少なくとも１つの合格ファイルがあったか否かを判定し（ステップＳ13）、１つもなければ、例えば入力された検索条件に合致する文書ファイルは１つもなかった旨を利用者に表示する等の処理を行って、処理を終了する。他方、１つでも合格ファイルが存在した場合、その全ての合格ファイルの文書ファイル名とキーワードフィルタ部５から出力された全ての属性名インデックスとを文書内容出力制御部10に通知して、文書内容出力制御を開始させる（ステップＳ14）。

【００５２】文書内容出力制御部10は、通知された１つの合格ファイル名に注目してその文書内容を取り出し、通知された全ての属性名インデックスとともに文書内容出力部９に渡し、当該文書の処理を開始させる（ステップＳ15）。

【００５３】文書内容出力部９は、通知された１つの属性名インデックスに注目し（ステップＳ16）、その属性名インデックスの属性名をもつ属性名タグが文書内にあるかを調べ（ステップＳ17）、あれば（ステップＳ18でＹＥＳ）、その属性名タグに対応する属性値タグの値と当該文書ファイル名とを表示装置１０２に表示する（ステップＳ19）。なければ（ステップＳ18でＮＯ）、ステップＳ19をスキップする。次に文書内容出力部９は、通知された属性名インデックスに未処理の属性名インデックスが残っているか否かを調べ（ステップＳ20）、残っていれば、その１つに注目を移し（ステップＳ21）、ステップＳ17に戻って上述した処理を繰り返す。

【００５４】文書内容出力部９が通知された全ての属性名インデックスについての処理を終えると（ステップＳ20でＮＯ）、文書内容出力制御部10は、合格文書ファイル名選別部８から通知された文書ファイルに未処理のものが残っているか否かを調べ（ステップＳ22）、残っている場合にはその１つに注目を移し、その文書ファイル名の文書内容を取り出して、合格文書ファイル名選別部８から通知された全ての属性名インデックスとともに文書内容出力部９に渡し、処理させる（ステップＳ23）。全ての合格ファイルについての処理が終わると（ステップＳ22でＹＥＳ）、処理終了となる。

【００５５】

【実施例】文書ファイル名辞書１に、図１に例示するように「ファイル１」，「ファイル２」，「ファイル３」の３つの文書ファイル名とそのＵＲＬとが登録されているとする。また、ファイル１の内容が図５（ａ）に示すものであり、ファイル２の内容が図５（ｂ）に示すものであるとする。これらのファイル１，２はXML で記述されており、文章テキストの他に属性と属性値が含まれている。つまり、ファイル１には、＜ホテル＞Ｘホテル＜／ホテル＞、＜場所＞サンフランシスコ＜／場所＞、＜値段＞＄１５０＜／値段＞、＜ペット＞可能＜／ペット＞といった、文書中に書かれた意味を表現する属性名のついたタグとその属性の値との対が含まれている。同様に、ファイル２にも、＜ホテル＞Ｚホテル＜／ホテル＞、＜場所＞シアトル＜／場所＞、＜値段＞＄１８０＜／値段＞、＜ペット＞不可＜／ペット＞といったタグが含まれている。

【００５６】また、属性名格納辞書２には図１に例示するような属性名とその自然言語表現との対が事前に格納されており、属性値格納辞書３には図１に例示するような属性値と自然言語表現と属性名との３つ組が事前に格納されているものとする。なお、属性値格納辞書３に全ての価格をその実際値で登録すると、登録数が増えてしまうので、変数を使用して登録するようにしても良い。つまり、XXX を任意の数値とする場合、以下の表３に示すように登録しておき、キーワード抽出部４は任意の数値の後ろに「ドル」があれば、自然言語表現XXX ドルが存在すると判断し、存在した実際値の頭に＄を付けたものを属性値とする。

【表３】

属性値	自然言語表現	属性名
$XXX	XXXドル	価格

【００５７】このような前提で、利用者が以下のような自然言語による検索入力文を入力した場合を例に、本実施例の動作を説明する。

検索入力文：「値段が150 ドルでペットを連れて行くのが可能で、しかも、サンフランシスコ近辺にあるホテルの情報を見つけたい」

【００５８】キーワード抽出部４は利用者からの検索入力文を受け付けると、属性名格納辞書２および属性値格納辞書３を参照して、検索入力文を以下のようにキーワード列に変換する。

【００５９】まず、検索入力文の先頭の自然言語表現「値段」が属性名格納辞書２に存在するので、それと対になって登録されている属性名「価格」を属性名インデックスとして出力する。次に、自然言語表現「150 ドル」が属性値格納辞書３に存在するので、それと３つ組で登録されている属性値「＄150 」と属性名「価格」との対を属性値インデックスとして出力する。次に、自然言語表現「ペット」が属性名格納辞書２に存在するので、それと対になって登録されている属性名「ペット」を属性名インデックスとして出力する。次に、自然言語表現「可能」が属性値格納辞書３に存在するので、それと３つ組で登録されている属性値「可能」と属性名「ペット」との対を属性値インデックスとして出力する。次に、自然言語表現「サンフランシスコ」が属性値格納辞書３に存在するので、それと３つ組で登録されている属性値「サンフランシスコ」と属性名「場所」との対を属性値インデックスとして出力する。次に、「ホテル」が属性名格納辞書２に存在するので、それと対になって登録されている属性名「ホテル」を属性名インデックスとして出力する。検索入力文中には、属性名格納辞書２および属性値格納辞書３に登録された自然言語表現とマッチする他の自然言語表現はない。従って、以下のようなキーワード列が上から順に出力される。

【００６０】属性名インデックス（属性名「価格」）
属性値インデックス（属性値「$150」、属性名「価格」）
属性名インデックス（属性名「ペット」）
属性値インデックス（属性値「可能」、属性名「ペット」）
属性値インデックス（属性値「「サンフランシスコ」、属性名「場所」）
属性名インデックス（属性名「ホテル」）

【００６１】次にキーワードフィルタ部５は、属性名インデックスと属性値インデックスとの並び順を参照して、冗長な部分の統合を行う。上のキーワード列の場合、属性名インデックス（属性名「価格」）と属性値インデックス（属性値「$150」、属性名「価格」）とは同じ属性名「価格」で隣どうしに並んでいるので、属性名インデックス（属性名「価格」）を削除する。また、属性名インデックス（属性名「ペット」）と属性値インデックス（属性値「可能」、属性名「ペット」）とは同じ属性名「ペット」で隣どうしに並んでいるので、属性名インデックス（属性名「ペット」）を削除する。他に削除すべき属性名インデックスは存在しないので、上記のキーワード列は最終的に以下のように圧縮される。

【００６２】(a) 属性値インデックス（属性値「$150」、属性名「価格」）
(b) 属性値インデックス（属性値「可能」、属性名「ペット」）
(c) 属性値インデックス（属性値「「サンフランシスコ」、属性名「場所」）
(d) 属性名インデックス（属性名「ホテル」）

【００６３】次に、合格文書ファイル名選別部８は、文書ファイル名辞書１中のファイル１の文書内容をそのＵＲＬを頼りにインターネット103 を通じて該当するサーバから取得し、その文書内容と上記の属性値インデックス(a) ～(c) とを統合文書内容検査部７に渡す。

【００６４】統合文書内容検査部７は、ファイル１の文書内容と、１つの属性値インデックス(a) とを文書内容検査部６に渡す。

【００６５】文書内容検査部６は、ファイル１の文書内容中に、属性値インデックス(a) 中の属性名「価格」のタグが存在するか否かを調べる。図５（ａ）のファイル１の場合、該当するタグ＜値段＞$150＜／値段＞があるので、その属性値「$150」が受け取った属性値インデックス(a) 中の属性値「$150」と一致するか否かを調べる。今の例では、一致するので、合格を統合文書内容検査部７に返却する。

【００６６】統合文書内容検査部７は、ファイル１の文書内容と、次の属性値インデックス(b) とを文書内容検査部６に渡す。

【００６７】文書内容検査部６は、ファイル１の文書内容中に、属性値インデックス(b) 中の属性名「ペット」のタグが存在するか否かを調べる。図５（ａ）のファイル１の場合、該当するタグ＜ペット＞可能＜／ペット＞があるので、その属性値「可能」が受け取った属性値インデックス(b) 中の属性値「可能」と一致するか否かを調べる。今の例では、一致するので、合格を統合文書内容検査部７に返却する。

【００６８】統合文書内容検査部７は、ファイル１の文書内容と、次の属性値インデックス(c) とを文書内容検査部６に渡す。

【００６９】文書内容検査部６は、ファイル１の文書内容中に、属性値インデックス(c) 中の属性名「場所」のタグが存在するか否かを調べる。図５（ａ）のファイル１の場合、該当するタグ＜場所＞サンフランシスコ＜／場所＞があるので、その属性値「サンフランシスコ」が受け取った属性値インデックス(c) 中の属性値「サンフランシスコ」と一致するか否かを調べる。今の例では、一致するので、合格を統合文書内容検査部７に返却する。

【００７０】統合文書内容検査部７は、ファイル１に関し全ての属性値インデックスで合格の結果が得られたので、合格文書ファイル名選別部８に合格を通知し、合格文書ファイル名選別部８はファイル１を合格ファイルとする。

【００７１】次に合格文書ファイル名選別部８は、文書ファイル名辞書１に格納されたファイル２の文書内容をそのＵＲＬを頼りにインターネット103 を通じて該当するサーバから取り込み、先のファイル１と同様に統合文書内容検査部７を使って合否を判定する。この場合、ペット属性、場所属性が満足しないので、ファイル２は不合格となる。同様に、残りのファイル３についても合否の判定が行われる。ここでは、ファイル３も不合格と判定され、合格ファイルはファイル１のみであったとする。

【００７２】次に合格文書ファイル名選別部８は、合格ファイル名としてファイル名１を、属性名インデックス(d) とともに文書内容出力制御部10に渡す。

【００７３】文書内容出力制御部10は、文書ファイル名辞書１からファイル名１のＵＲＬを取得し、それを頼りにインターネット103 上のサーバをアクセスしてファイル名１の文書内容を取得し、属性名インデックス(d) とともに文書内容出力部９に渡す。

【００７４】文書内容出力部９は、ファイル１の文書内容中に、属性名インデックス(d) の属性名「ホテル」を持つ属性タグが存在するか否かを調べる。図５（ａ）のファイル１の場合、該当するタグ＜ホテル＞Ｘホテル＜／ホテル＞が存在するので、その属性値「Ｘホテル」を取り出し、ファイル名１と共に表示装置102 に表示する。

【００７５】図４は本発明の文書ファイル検索装置のハードウェア構成例を示すブロック図である。この例の文書ファイル検索装置は、プロセッサ（ＣＰＵ）２００と、磁気ディスク装置等の補助記憶装置２０１と、そのインタフェース２０２と、ＲＡＭ等のメモリ２０３と、インターネット１０３との間のインタフェース２０４と、ＣＤ－ＲＯＭ，半導体メモリ等の機械読み取り可能な記録媒体２０５と、そのインタフェース２０６と、入力装置１０１と、そのインタフェース２０７と、表示装置１０２と、そのインタフェース２０８と、ＣＰＵ２００，メモリ２０３およびインタフェース２０２，２０

４，２０６～２０８間を接続するバス２０９とから構成されている。

【００７６】記録媒体２０５には、文書ファイル検索用プログラムが記録されており、このプログラムがインタフェース２０６を介してインストールされることにより、メモリ２０３または補助記憶装置２０１上に図１の文書ファイル名辞書１，属性名格納辞書２および属性値格納辞書３がロードされる。また、同プログラムはＣＰＵ２００の動作を制御することにより、ＣＰＵ２００を図１のキーワード抽出部４、キーワードフィルタ部５、文書内容検査部６、統合文書内容検査部７、合格文書ファイル名選別部８、文書内容出力部９、文書内容出力制御部10として機能させる。

【００７７】以上の実施の形態は本発明をWWW 上のホームページの検索に適用したが、特開平５－６７１３６号公報に記載する技術と同様にデータベースに対する検索に対しても適用可能である。

【００７８】

【発明の効果】以上説明したように本発明によれば以下のような効果が得られる。

【００７９】自然言語による検索問い合わせに対する回答の冗長性を極力無くすことができる。その理由は、キーワード抽出部で抽出されたインデックス列をその先頭から順に探査し、同一の属性の属性名インデックスと属性値インデックスとが隣どうしに存在する場合に属性名インデックスを削除するキーワードフィルタ部を備えているからである。

【００８０】WWW のホームページに対しても自然言語による検索問い合わせが可能になる。その理由は、XML のように文書ファイル中にその意味を表現する属性名と属性値のタグを内蔵させており、利用者が入力した自然言語による検索条件を解釈して適合する属性名および属性値をもつ文書ファイル中から、利用者の望む属性値を取り出すことができるからである。これによって、文法的に正しい自然言語表現、非文法的な表現、自然言語文の断片、キーワード列等、種々の形の入力を受け付けて統一的に解釈を行なう自然言語インタフェースによるWWW文書検索システムを実現することが出来る。

【図面の簡単な説明】

【図１】本発明の実施の形態の文書ファイル検索装置のブロック図である。

【図２】本発明の実施の形態の文書ファイル検索装置処理例を示すフローチャートである。

【図３】本発明の実施の形態の文書ファイル検索装置処理例を示すフローチャートである。

【図４】本発明の文書ファイル検索装置のハードウェア構成例を示すブロック図である。

【図５】XML を使った文書ファイルの記述例を示す図である。

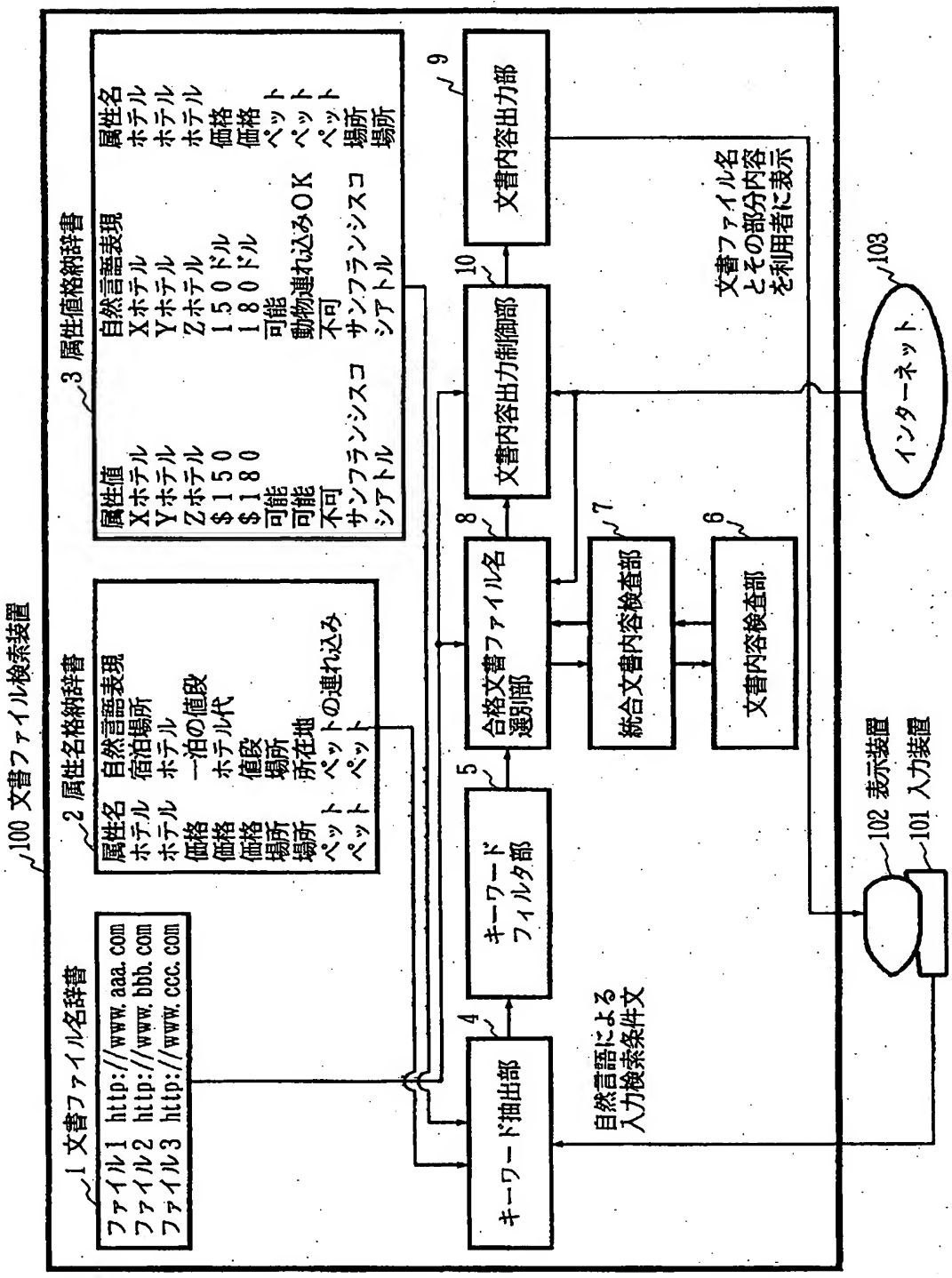
【図６】HTMLの記述例を示す図である。

【図７】図６のHTMLの記述例をブラウザで表示した例を示す図である。

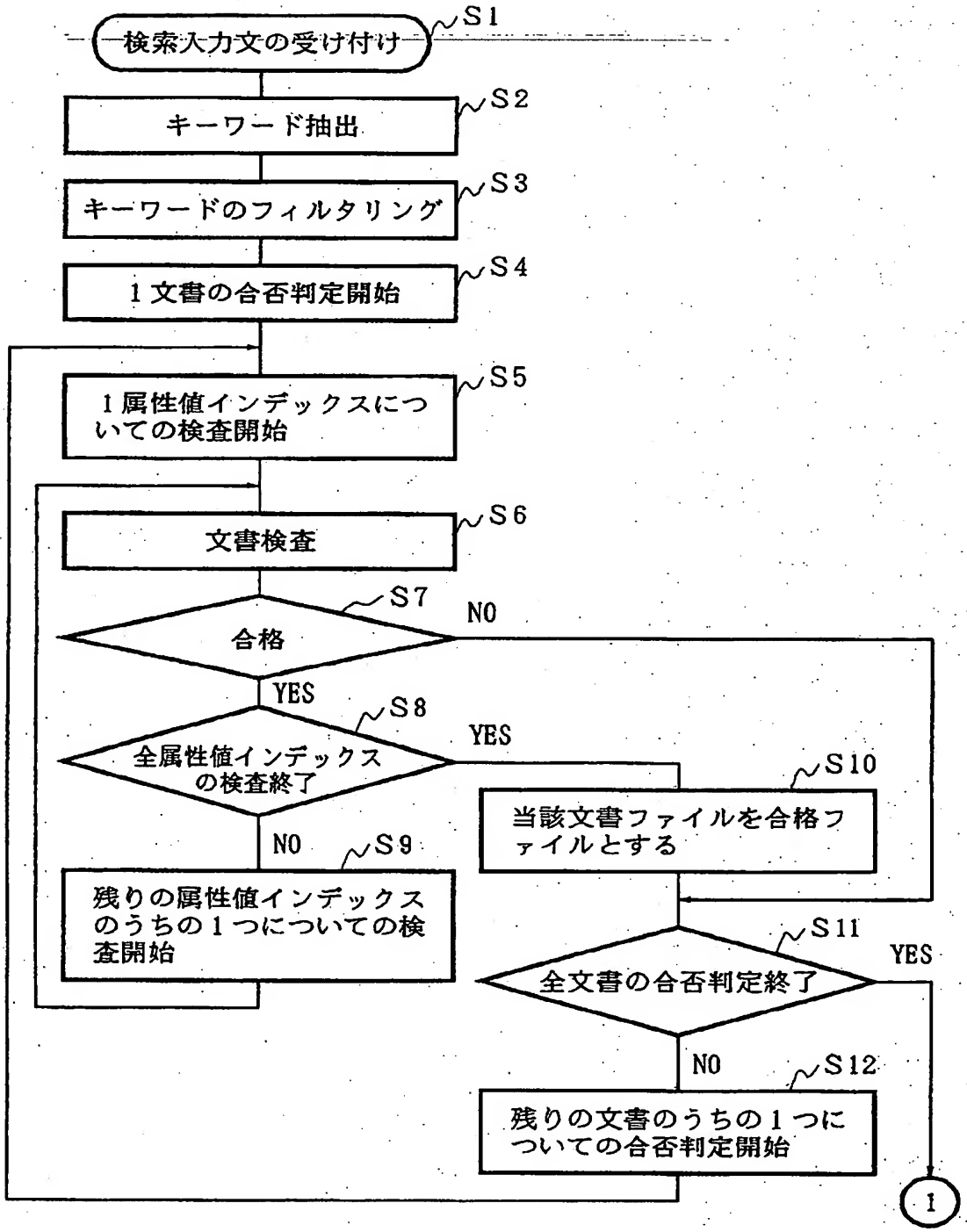
【符号の説明】

１は文書ファイル名辞書、２は属性名格納辞書、３は属性値格納辞書、４はキーワード抽出部、５はキーワードフィルタ部、６は文書内容検査部、７は統合文書内容検査部、８は合格文書ファイル名選別部、９は文書内容出力部、10は文書内容出力制御部、である。

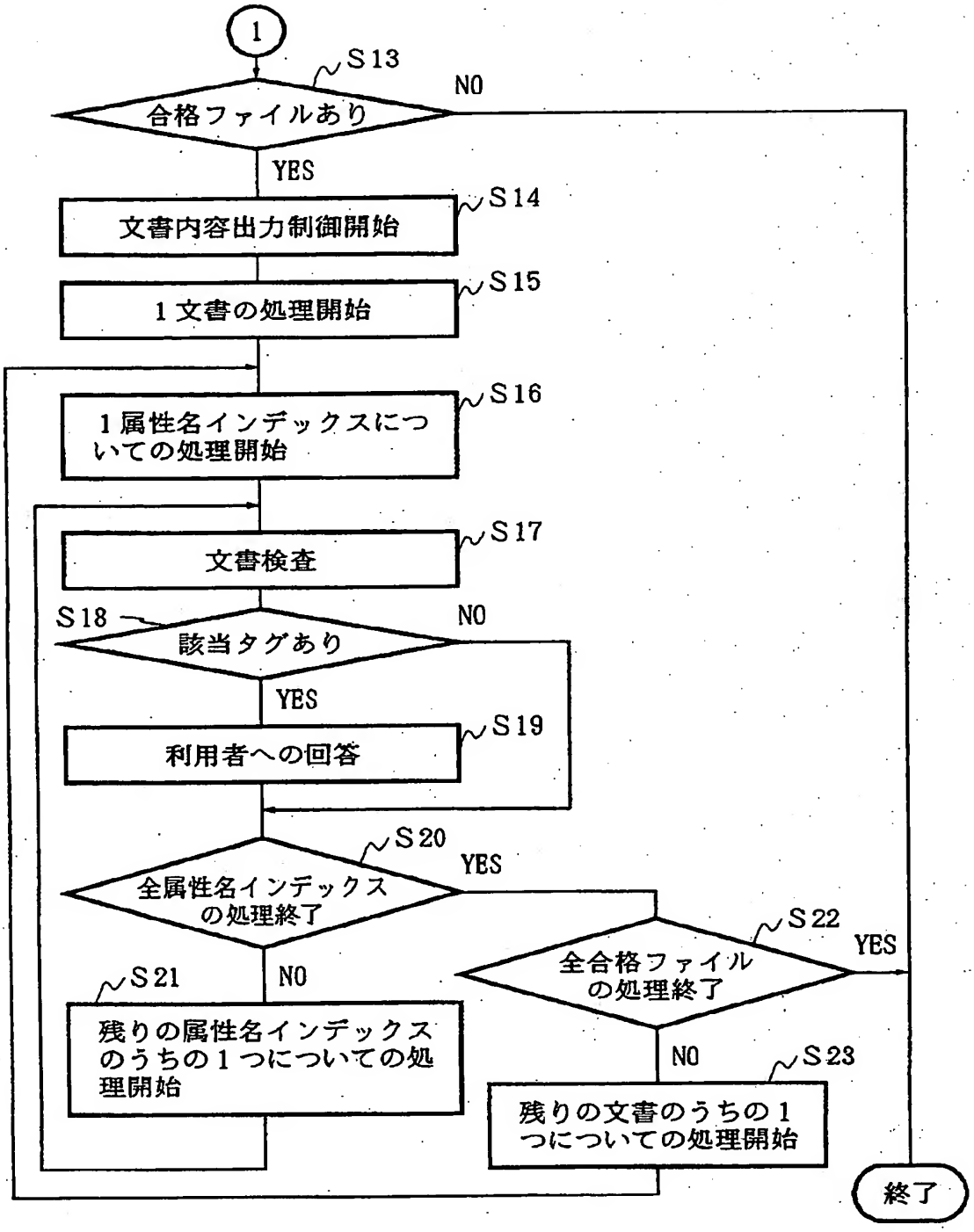
【図１】


100 文書ファイル検索装置
1 文書ファイル名辞書
ファイル1 http://www.aaa.com
ファイル2 http://www.bbb.com
ファイル3 http://www.ccc.com
2 属性名格納辞書
属性名 自然言語表現
ホテル 宿泊場所
ホテル ホテル
価格 一泊の値段
価格 ホテル代
価格 値段
場所 場所
場所 所在地
ペット ペットの連れ込み
ペット ペット
3 属性値格納辞書
属性値 自然言語表現 属性名
Xホテル Xホテル ホテル
Yホテル Yホテル ホテル
Zホテル Zホテル ホテル
＄１５０ １５０ドル 価格
＄１８０ １８０ドル 価格
可能 可能 ペット
可能 動物連れ込みOK ペット
不可 不可 ペット
サンフランシスコ サンフランシスコ 場所
シアトル シアトル 場所
4 キーワード抽出部
5 キーワード フィルタ部
6 文書内容検査部
7 統合文書内容検査部
8 合格文書ファイル名 選別部
9 文書内容出力部
10 文書内容出力制御部
自然言語による 入力検索条件文
文書ファイル名 とその部分内容 を利用者に表示
インターネット
103
102 表示装置
101 入力装置

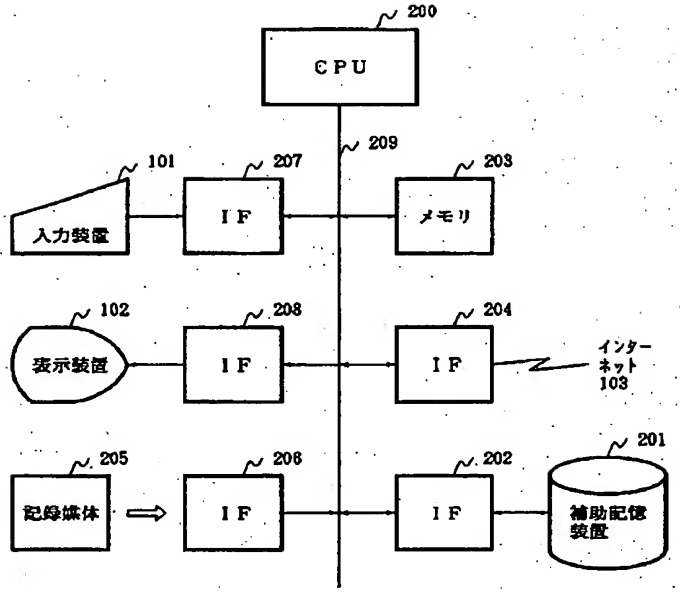
【図２】


検索入力文の受け付け
S1
キーワード抽出
S2
キーワードのフィルタリング
S3
１文書の合否判定開始
S4
１属性値インデックスについての検査開始
S5
文書検査
S6
合格
S7
NO
YES
全属性値インデックスの検査終了
S8
YES
NO
残りの属性値インデックスのうちの１つについての検査開始
S9
当該文書ファイルを合格ファイルとする
S10
全文書の合否判定終了
S11
YES
NO
残りの文書のうちの１つについての合否判定開始
S12
1

【図３】


1
S13
合格ファイルあり
NO
YES
S14
文書内容出力制御開始
S15
１文書の処理開始
S16
１属性名インデックスについての処理開始
S17
文書検査
S18
該当タグあり
NO
YES
S19
利用者への回答
S20
全属性名インデックスの処理終了
YES
S22
全合格ファイルの処理終了
YES
S21
NO
残りの属性名インデックスのうちの１つについての処理開始
NO
S23
残りの文書のうちの１つについての処理開始
終了

【図４】


200
CPU
209
101
207
203
入力装置
IF
メモリ
102
208
204
表示装置
IF
IF
インターネット
103
205
206
202
201
記録媒体
IF
IF
補助記憶装置

【図5】

(a)

```
<!--Xホテルのプロフィールのファイル-->  ⟸これはコメント行
<?xml version="1.0",encoding="shiftjis"?>
<!DOCTYPE 宿泊情報プロフィールSYSTEM"http://report/hotel.xml">
<プロフィール記録>
<ホテル>Xホテル</ホテル>
<場所>サンフランシスコ</場所>
<値段>$150</値段>
<ペット>可能</ペット>

<プロフィール>
ホテルXは、ビジネス街に位置し、ビジネス、観光いずれにも、・・・
</プロフィール>
・・・
</プロフィール記録>
```

(b)

```
<!--Zホテルのプロフィールのファイル-->  ⟸これはコメント行
<?xml version="1.0",encoding="shiftjis"?>
<!DOCTYPE 宿泊情報プロフィールSYSTEM"http://report/hotel.xml">
<プロフィール記録>
<ホテル>Zホテル</ホテル>
<場所>シアトル</場所>
<値段>$180</値段>
<ペット>不可</ペット>

<プロフィール>
ウォータフロントの中心部にあるZホテルは、シアトルで最もホットな場所として、・・・
</プロフィール>
・・・
</プロフィール記録>
```

【図６】

```
<html><title>1201</title>
<body BGCOLOR="#eeeeee" TEXT="#000000">
<TABLE BORDER="1" CELLSPACING="1" WIDTH="550">
<TR><TH>担当課題名</TH><TD COLSPAN="2">ＸＭＬの自然言語インタフェース方式の開発</TD></TR>
<TR><TH>担当名</TH><TD>ＮＥＣ　Ｃ&Ｃメディア研究所</TD></TR>
<TR><TH>優先度</TH><TD>高い</TD></TR>
</TABLE>
<P>
<B>１．研究目的</B><P>
本研究は、先端的であり、・・・・
<P>
<B>２．具体的既往成果</B><P>
本研究部門では、すでに、自然言語インタフェースの研究を、・・・・
<P>
<B>３．全体計画</B><P>
１年目には、・・・・、２年目には、・・・・
</body></HTML>
```

【図７】

担当課題名　　ＸＭＬの自然言語インタフェース方式の開発
担当名　　ＮＥＣ　Ｃ＆Ｃメディア研究所
優先度　　高い

１．研究目的

本研究は、先端的であり、・・・・

２．具体的既往成果

本研究部門では、すでに、自然言語インタフェースの研究を、・・・・

３．全体計画

１年目には、・・・・、２年目には、・・・・